Psychometric Review of Language Tests for Preschool Children

A Senior Honors Thesis

Presented in Partial Fulfillment of the Requirements for graduation *with research distinction* in Speech and Hearing Science in the undergraduate colleges of The Ohio State University

by

Abby Kinsey

The Ohio State University

Project Advisor: Dr. Rebecca McCauley, Department of Speech and Hearing Science

Abstract

Speech-language pathologists have a large number of norm-referenced tests to choose from as they diagnose language disorders in preschool age children. Because determining the psychometric quality of a test is important to that choice, the purpose of the present study is to review the quality of currently available tests. A standardized search strategy yielded 15 norm-referenced tests published since 1998 that focused on language skills in preschool children. Eleven criteria related to the test manual's documentation of evidence such as reliability and validity were developed for use in the review based on those used in 2 similar studies (McCauley & Swisher, 1984; McCauley and Strand, 2008). Studies such as this can (a) promote ongoing study of existing tests to strengthen evidence of their psychometric quality and (b) encourage improvements in future test development. Further, such studies can increase speech-language pathologists' attention to the importance of psychometric characteristics and the potentially negative effects of poorer tests on the quality of their own work.

Table of Contents

Acknowledgements

I would like to thank my advisor, Dr. Rebecca McCauley for allowing me to work alongside her on this project. Working alongside her I have grown professionally in the field of speech language pathology as well as in the area of research. I would also like to thank Pearson and Pro-Ed for donating many of the tests used in the study. Finally, I would like to thank my friends and family for all the support they have given me throughout this process.

Introduction

This study is a replication of a 1984 study by Rebecca McCauley and Linda Swisher. The previous study examined both language and articulation tests. This particular study replicates the language portion of the previous study.

There are 3 appropriate assessment objectives for which tests are used to (a) determine the existence of a speech or language disorder, (b) determine the goals of intervention, and (c) plan procedures for intervention. Norm-referenced tests are primarily used to determine the existence of a disorder. That is because norm-referenced tests measure the child's ability compared to the normative scores of the sample of children in the child's age range. If the child's score falls a predetermined distance (e.g., 1.5 SD) below the mean score of the normative sample, then it is said that he/she has may have a speech or language disorder.

The two purposes for this study are (a) to stimulate discussion of the psychometric characteristics of language tests rather than to serve as a definitive psychometric review and (b) to see how psychometric data has changed for tests in the last 25 years. Therefore, the criteria used in the review focused on a selected sample of a larger number of important psychometric characteristics. They included those used in the previous study (McCauley & Swisher, 1984) as well as an additional one that was added to more fully examine validity evidence and was similar to one used in a more recent psychometric review (McCauley & Strand, 2008).

There are three concepts that are important to understand and take into consideration in order to have a true understanding of norm-referenced tests and the standardized process. These concepts are (a) test validity and reliability, (b) the normative sample, and (c) test norms and derived scores.

Background on basic psychometric concepts

Validity and Reliability

A measurement instrument, such as a language test, is described as psychometrically valid if it accurately measures what it is designed to measure. There are four types of validity evidence that can help demonstrate the successful construction of a test. The four types of evidence are evidence of content validity, concurrent validity, predictive validity, and construct validity. Each of these individual characteristics should be considered and present in every language test manual.

Content validity procedures involve the systematic examination of the test content to determine whether it covers a representative sample of the behavior domain to be measured (Anastasi & Urbina, 1997). One form of criterion-related validity is termed *concurrent validity*. Evidence for this type of validity should indicate the effectiveness of a test in determining an individual's performance in specific activities (Anastasi & Urbina, 1997). If a language test shows validity in this area then the test should correlate well with other tests that are measuring the same ability. *Predictive validity*, the other form of evidence usually considered under the label *criterion-related validity is* examined by assessing how closely an individual's test score can be used to predict future performance on a criterion measure. The information that would be supplied by a test validated in this fashion could provide an important basis for the identification of subgroups differentiated by prognosis (McCauley & Swisher, 1984).*Construct validity* of a test is the extent to which the test may be said to measure a theoretical construct or trait. (Anastasi & Urbina, 1997, p. 126) This type of validity is studied by a careful comparison of the test author's description of the construct to be tested to the test's actual content. In order for a test to measure

what it is intended to measure, evidence that there is high reliability must also be reported in a test manual. Reliability refers to the stability with which a test measures a given behavior. There are two types of reliability that should be present in a language test manual to ensure that it is a stable measurement. *Test-retest reliability* is the first in which the manual for a language test should include. This should examine the extent to which a test taker's performance is consistent over a period of time. This type of reliability shows the extent to which scores on a test can be generalized over different occasions; the higher the reliability, the less susceptible the scores are to random daily changes in the conditions of the examinee or the testing environment (Anastasi & Urnina, 1997). *Interexaminer reliability* is the second type of reliability that should be provided in the manual for a test. Evidence of this type of reliability is provided through a study demonstrating a correlation between two scorers. This type of reliability is used to show that the directions and administration procedures are clear enough so that multiple scorers will yield approximately the same score.

The Normative Sample

Test norms are "a statistical summary of the scores received by a normative sample" (McCauley & Swisher, 1984, p.36). These norms are a basis for comparing a child's score with peers based on age and language experience. In order to find the existence of a problem, it may be appropriate to obtain norms from different normative groups. Test developers should ideally publish norms for a variety of different groups to allow the test user to answer all relevant assessment questions for all possible test takers. However, it is rare that a representative normative sample is provided for even the most basic question—Is there a language impairment?

When a norm-referenced language test is developed, the normative sample should clearly be defined in the test manual. The information that should be provided to clearly define the normative sample include age, geographic residence, and socioeconomic status. These factors are important to know because they can have an effect on the child's performance, therefore having an effect on how the scores should be interpreted. Age would have an effect because you would expect that a 10 year old child would do better on the test than a 3 year old child. Geographic residence is important as well because it is possible that there would be dialectal differences in individuals which could potentially affect the way in which a child scores on the test. And lastly, socioeconomic status is important to know because studies have shown that lower socioeconomic status can be associated with poorer test performance on existing language tests (Arnold & Reed, 1976; Johnson, 1974).

Along with the description of the normative sample, it is also important for the test user to know how the sample was chosen. Often, a test designer will exclude individuals who have a disability or non-normal language abilities. According to McCauley and Swisher (1984), this can present difficulties because even the most deviant scores within the normative sample are still representing normal performance. This makes it hard to tell whether a child that scores just below the lowest score has nonnormal performance which makes it difficult to tell just how different the score has to be before it reflects a language impairment.

Test Norms and Derived Scores

Norm-referenced tests should provide the mean and standard deviation for the total raw scores for all relevant subgroups. The means should be provided because it is a factor that

determines, based off the normative sample, the score range in which the child should be testing based off of his/her age. The standard deviations should be provided because it allows the test user to determine whether a child has a language disorder based off of how far away a child's score is from the average. It is important when looking at the means and standard deviations to also look at the derived scores that correlate with and assist in finding the means and standard deviations.

Norm-referenced tests typically have three types of derived scores that are used for interpretation of test norms and test taker's scores. These three scores are developmental or age equivalent scores, percentiles, and standard scores. These scores will be briefly described for why they are appropriate. Age equivalent or developmental scores are included because they present the test norms in a way that shows how behaviors or abilities change with age. Percentile rank scores describe the percentage of those in the normative sample whose test scores fell below a given value. Since the percentiles indicate the position of a test taker's score compared to the scores of the normative sample, this type of score makes it easy to see how the test taker compares to the normative sample. The standard score calculations use information about the average score and variability of scores obtained by the normative sample. These scores can be used to estimate the position of a test taker's score relative to the scores obtained by the normative sample, to compare one's score on two different tests, and to compare one person's score to someone else's in a meaningful way (McCauley & Swisher, 1984, p.37).

Method

Tests to be reviewed were found through the BurosOnline Review and appropriate areas of the website associated with the American Speech-Language Hearing Association. Tests were excluded from the final list if they were: multiples, not norm-referenced, published prior to 1998, screenings, not specific for language disorders, or were not appropriate for preschool children. In the end, 15 tests were appropriate to use in the study. I examined each language test using the 11 criteria described below, with the outcome of my evaluation for each criterion indicated on a record sheet.

In order to obtain reliability about the coding for met and unmet criteria, a second examiner rated 20% of the tests in the review. The reviewer was an undergraduate colleague conducting a related portion of the McCauley and Swisher (1984) replications. Both of us had the same background knowledge and training in psychometric theory and criteria coding. In order to train, my colleague and I each examined 3 adult language tests and then compared out results.

The 11 criteria used in the review were chosen because of their importance and significance to tests of language, and because they could be translated into relatively objective decision rules. If the test does not meet these criteria or fails to provide information on them, we believe that the test user should have serious qualms about the quality of the test—no matter how familiar the test user is with the content of the test and the skills being tested. The 11 criteria used in this study are listed below; Criteria 1-10 appear just as they did in the 1984 study by

McCauley and Swisher and criterion 11, which used to evaluate construct validity, was adapted from a criterion used in McCauley and Strand (2008).

Criterion 1. The test manual should clearly define the standardization sample so that the test user can examine its appropriateness for a particular test taker (APA, 1974; Weiner & Hoock, 1973). In order to pass this criterion, the test manual needed to give three pieces of information considered important for speech and language testing. : (a) the normative sample's geographic residence, (b) socioeconomic status, and (c) the "normalcy" of subjects in the sample, including the number of individuals excluded because they exhibited nonnormal language or nonnormal, general development (Salvia & Ysseldyke, 1981).

Consequences if unmet. If the test manual does not have this information, the test user cannot tell whether the normative sample is representative of the test author's intended population. If the test user does not know the intended population, then the test user does not know if it is a population that the test taker's performance should be compared to.

Criterion 2. For each subgroup examined during the standardization of the test, an adequate sample size should be used (APA, 1974). In order for the test to pass this criterion, the test needed to have subgroups with sample sizes of 100 or more. This particular value is consistently referred to by the authorities as the lower limit for adequate sample sizes (Salvia & Ysseldyke, 1981; Weiner & Hook, 1973).

Consequences if unmet. If a small sample is used then the norms are likely to be less reliable and stable. If the sample is small, it is likely that a different group of children might have different norms. Also, a small sample may not have relatively rare individuals such as those with language or articulation disorders, making it difficult to interpret the scores of those individuals that are possibly impaired.

Criterion 3. The reliability and validity of the test should be promoted through the use of systematic item analysis during item construction and selection (Anastasi, 1976, p. 198). To pass this criterion, the test manual needed to report evidence that quantitative methods were used to study and control item difficulty, item validity, or both.

Consequences if unmet. Without this information, the test user is unaware as to whether the test accurately measures what it purports to measure. This criterion is to make sure the test possesses validity and reliability.

Criterion 4. Measures of the central tendency and variability of test scores should be reported in the manual for relevant subgroups examined during the objective evaluation of the test (APA, 1974, p. 22). For the test to pass this criterion, both the mean and standard deviation had to be given for the total raw scores of all relevant subgroups.

Consequences if unmet. The mean gives the average score of the normative sample's relevant subgroups. The standard deviation gives the test user and estimate how much variation there was in scores received by the individuals in

the subgroups. This information serves as the basis for other ways of presenting norms (e.g., z scores). The absence of this information causes the test user to lose flexibility in the use of the test norms.

Criterion 5. Evidence of concurrent validity should be supplied in the test manual (APA, 1974, pp. 26-27).To pass this criterion, the test manual needed to provide empirical evidence that categorizations of children as normal or impaired obtained using the test agree closely with categorizations obtained by other methods that can be considered valid, for example, clinician judgments or scores on other validated tests.

Consequences if unmet. Without this type of reliability, it questions a test's ability to correctly answer assessment questions that relate to the existence of a language impairment. Failing this criterion should cause the clinician to question the usefulness of the test because the reasoning for using norm-referenced tests of language is to allow the test user to compare a child's score with other children to determine normalcy.

Criterion 6. Evidence of predictive validity should be supplied in the test manual (APA, 1974, pp. 26-27).In order to pass this criterion, a test manual needed to include empirical evidence that it could be used to predict later performance on another, valid criterion of the speech or language behavior addressed by the same test in question.

Consequences if unmet. This type of validity is used in order to make assessment decisions by a clinician related to the need for therapy. Absence of this validity

means that it is possible for invalid sources of information to be weighted more heavily in a decision process.

Criterion 7. An estimate of test-retest reliability for relevant subgroups should be supplied in the test manual (APA, 1974, pp. 50, 54). To pass this criterion, the test manual needed to supply empirical evidence of test-retest reliability, including a correlation coefficient of .90 or better (Salvia & Ysseldyke, 1981, p. 98) that was statistically significant at or beyond the .05 level (Anastasi, 1976, pp. 108-109)

Consequences if unmet. Without this type of reliability, it questions the extent in which the test results are stable and how it may fluctuate overtime. Children's test-retest reliability should have a minimum correlation coefficient of .90 and be statistically significant at .05 level or beyond. A correlation coefficient at this level would show that results from re-testing a child would be comparable in score with the first testing.

Criterion 8. Empirical evidence of interexaminer reliability should be given in the test manual (APA, 1974, p.50). To pass this criterion, a test manual needed to report evidence of interexaminer reliability that included a correlation coefficient of .90 or better (Salvia & Ysseldyke, 1981, p. 98) that was statistically significant at or beyond the .05 level (Anastasi, 1976, pp. 108-109).

Consequences if unmet. Without this reliability, the test user does not know the degree to which a test taker is likely to score similar if the test is given by

different clinician. It would be unknown whether the test taker is likely to affect the scores of test takers in a way that that could benefit or penalized them.

Criterion 9. Test administration procedures should be described in sufficient detail to enable the test user to duplicate the administration and scoring procedures used during test standardization (APA, 1974, p. 18). In order to pass this criterion, the test manual needed to provide sufficient description so that, after reading the test manual, the reviewer believed she could administer and score the test without grave doubts about correct procedures.

Consequences if unmet. Without a clear description of the administration procedures the test user does not know if it is reasonable to compare the scores to the norms. If the test procedures are not duplicated in the same way they were for the standardization, it could result in unfair advantage or disadvantage for the test taker.

Criterion 10. The test manual should supply information about the special qualifications required of the test administrator or scorer (APA, 1974, p.15; Salvia & Ysseldyke, 1981, p.18). To pass this criterion, the test manual needed to state both general and specialized training required for administrators and scorers.

Consequences if unmet. This information should be provided because the test should only be given by someone who is qualified. A qualified individual will have background knowledge and training in administration, scoring, and

interpretation of test results. If the test manual does not provide this information, it could call question to the quality of the data obtained within the test.

Criterion 11. Construct validity should be present. (Operational definition based on McCauley & Strand, 2008, p.84). Any of the following were needed to meet the operational definition: (a) evidence from a factor analytic study confirming expectations of the test's internal structure, (b) evidence that test performance improves with age, (c) evidence that groups that were predicted to differ in test performance actually do so. In addition, evidence needed to be obtained within a study in which statistical methods were described and participants were described.

Consequences if unmet. This information provides the test user with validity that proves the test correlates with the theoretical construct of the test, which allows the test user to be confident in the fact that the test is measuring what it purports to measure. Without this validity, the test user may question whether or not the test will accurately measure what they are trying to measure.

Results

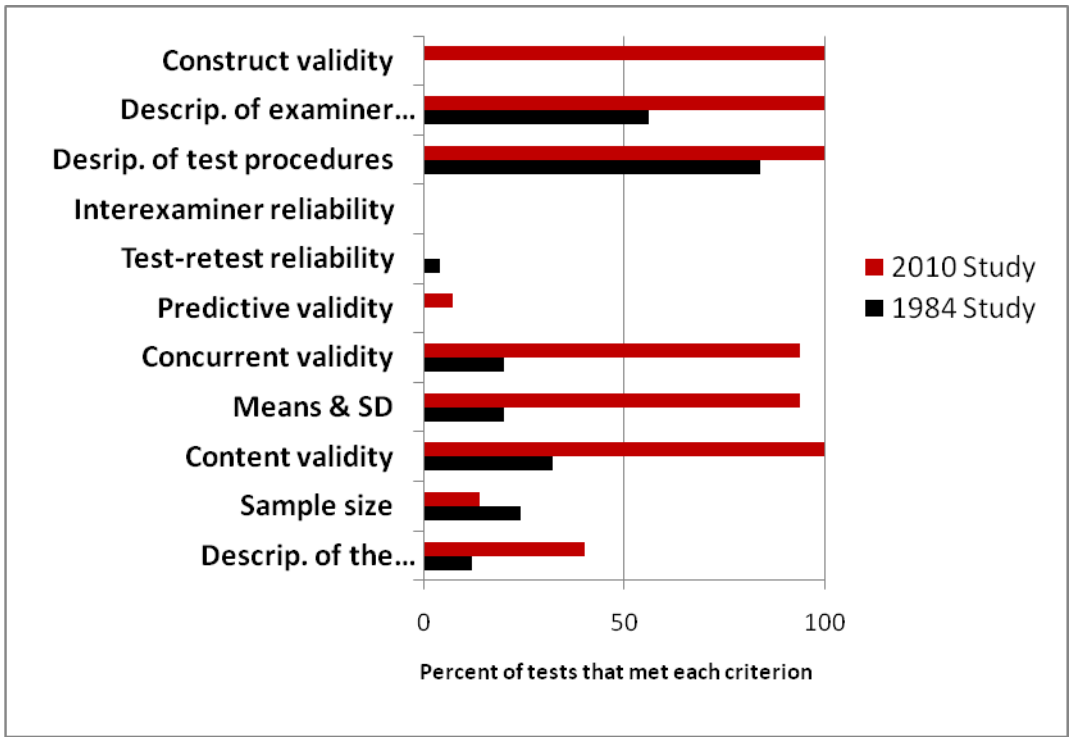
When we initially searched for tests appropriate for diagnosing and assessing language disorders, 439 tests were indentified in the search. Four hundred and twenty four tests were excluded from evaluation because they were multiples, screenings, not language tests, published before 1998, or not appropriate for preschool children. Thus 15 language tests were considered suitable for review. When 4 randomly selected tests were re-coded by the second examiner, the two examiners' ratings agreed 98% of the time, that is, on 43 of 44 rating judgments (4 tests x 11 criteria).

Table 1 shows that there are certain criteria that almost every test met in current language tests used. Description of examiner qualifications, description of test procedures, construct validity, and content validity were all met by 100% of the tests in the study. The table also shows that there were certain criteria that were not being met by the majority of language tests being used. Two criteria that were not met by a single test was test-retest reliability and interexaminer reliability. Predictive validity as well as sample size had a low number of tests meeting those criteria.

When examining the results of the criteria that were met by the tests in the study it is also important to see how the percentages compare to the previous study done in 1984. These numbers show how tests have become more reliable and valid within the last 25 years.

Figure 1

Shows how the results of the 1984 study compare to the present study.



The figure shows that in the present study a higher percent of the criteria were met than in the 1984 study. Test-retest reliability was only one criterion where the previous study had a higher percentage, although the previous study only had 4% (1 test) that met the criterion. The present and the previous study both had 0% of tests meeting interexaminer reliability. In each of the remaining categories, the present study had a 7% or higher increase from the original study in 1984.

Table 1

Shows the criteria that each test met.

Criterion	***Number of Tests***	***Tests***
1. Description of the Standard Sample	6	CREVT-2, PLAI-2, REEL-3, TNL, TWF-2, UTLD-4
2. Sample Size	2	CELF-4, SPELT-P2
3. Content Validity	15	CELF-4, CREVT-2, EOWPVT, EVT-2, PLAI-2, PLSI, PPVT-4, REEL-3, ROWPVT, SPELT-3, SPELT-P2, TACL-3, TNL, TWF-2, UTLD-4
4. Means and Standard Deviations	14	CELF-4, CREVT-2, EOWPVT, EVT-2, PLAI-2, PPVT-4, REEL-3, ROWPVT, SPELT-3, SPELT-P2, TACL-3, TNL, TWF-2, UTLD-4
5. Concurrent Validity	14	CELF-4, CREVT-2, EOWPVT, EVT-2, PLSI, PPVT-4, REEL-3, ROWPVT, SPELT-3, SPELT-P2, TACL-3, TNL, TWF-2, UTLD-4
6. Predictive Validity	1	TWF-2
7. Test-retest Reliability	0	
8. Interexaminer Reliability	0	
9. Description of Test Procedures	15	CELF-4, CREVT-2, EOWPVT, EVT-2, PLAI-2, PLSI, PPVT-4, REEL-3, ROWPVT, SPELT-3, SPELT-P2, TACL-3, TNL, TWF-2, UTLD-4
10. Description of Examiner Qualifications	15	CELF-4, CREVT-2, EOWPVT, EVT-2, PLAI-2, PLSI, PPVT-4, REEL-3, ROWPVT, SPELT-3, SPELT-P2, TACL-3, TNL, TWF-2, UTLD-4
11. Construct Validity	15	CELF-4, CREVT-2, EOWPVT, EVT-2, PLAI-2, PLSI, PPVT-4, REEL-3, ROWPVT, SPELT-3, SPELT-P2, TACL-3, TNL, TWF-2, UTLD-4

Conclusion

In this study, I conducted a review of 15 norm-referenced language tests designed for use with preschool children. Through the review, I found that as a group, currently available standardized norm-referenced tests do not meet many of the criteria we used to examine their psychometric quality. This conclusion is actually similar to that of the original study completed in 1984. In this study, the failures of tests to meet individual criteria generally resulted from the absence of required information, not from poor reported performance on the test. Consequently when certain psychometric criteria are left unmentioned in the test manual, test users are left to wonder whether the test is reliable or unreliable for their testing purposes. This outcome strongly suggests the need for greater disclosure and sometimes collection of data concerning important psychometric characteristics when the tests are normed and published. Further, it suggests that clinicians continue to need to be aware of psychometric principles as well as of the psychometric flaws that can be present in a test they decide to use. Becoming aware of the psychometric principles and flaws of the tests they consider using will allow clinicians to reduce the impact of those flaws on clinical decisions as well as stress that decisions can never be based on the results of tests alone.

References

American Psychological Association. (1974). *Standards for educational and psychological tests.* Washington,DC: APA.

American Speech-Language-Hearing Association. (2009). Directory of Speech-Language Pathology Assessment Instruments.

Retrieved November 14, 2009, from www.asha.org/SLP/assessment/.

Anastasi, A. (1976). *Psychological testing* (4th ed.). New York: Macmillan.

Arnold, K., & Reed, L. (1976). The Grammatic Closure subtest of the ITPA: A comparative study of black and white children. *Journal of Speech and Hearing Disorders, 4I,* 477-485

Buros Institute. (2009). Test Reviews Online. Retrieved October, 16, 2009, from www.unl.edu/buros.

Johnson, D. (1974). The influences of social class and race on language test performance and spontaneous speech of preschool children. *Child Development, 45,* 517-521.

McCauley, R., & Strand, E. (2008). A review of standardized tests of nonverbal oral and speech motor performance in children. *American Journal of Speech-Language Pathology*, *17*, 81-91.

McCauley, R., & Swisher, L. (1984). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders*, *49*, 34-42.

Salvia, J., & Ysseldyke, J. (1981). *Assessment in special and remedial education* (2nd ed.). Boston: Houghton Mifflin.

Appendix A

15 Language Tests Examined During Test Review

Blank, M., Rose, S., & Berlin, L.(2003). *Preschool Language Assessment Instrument, Second Edition* (PLAI-2). Austin, TX: Pro-Ed.

Brownell, R. (2000). *Expressive One-Word Picture Vocabulary Test, Third Edition* (EOWPVT). Novato, CA: Academic Therapy Publications.

Brownell, R. (2000). *Receptive One-Word Picture Vocabulary Test, Second Edition* (ROWPVT). Novato, CA: Academic Therapy Publications.

Bzoch, K., League, R., & Brown, V. (2003) *Receptive-Expressive Emergent Language Test, Third Edition* (REEL-3). Austin, TX: Pro-Ed

Carrow-Woolfolk, E. (1999) *Test of Auditory Comprehension of Language, Third Edition* (TACL-3). Austin, TX: Pro-Ed.

Dawson, J., Eyer, J., Fonkalsrud,, J., Stout, C., Tattersall, P., & Croley, K. (2005). *Structured Photographic Expressive Language Test-Preschool, Second Edition* (SPELT-P2). DeKalb, IL: Janelle Publications.

Dawson, J., Stout, C., & Eyer, J. (2003). *Structured Photographic Expressive Language Test, Third Edition* (SPELT-3). DeKalb, IL: Janelle Publications.

Dunn, L., & Dunn, D. (2007) *Peabody Picture Vocabulary Test, Fourth Edition* (PPVT-4). Minneapolis, MN: NCS Pearson Psychcorp.

German, D. (2000). *Test of Word Finding, Second Edition* (TWF-2). Austin, TX: Pro-Ed.

Gilliam, J., & Miller, L. (2006). *Pragmatic Language Skills Inventory* (PLSI). Austin, TX: Pro-Ed.

Gillam, R., & Pearson, N. (2004) *Test of Narrative Language* (TNL). Austin, TX: Pro-Ed.

Mecham, M. (2003) *Utah Test of Language Development, Fourth Edition* (UTLD-4). Austin, TX: Pro-Ed.

Semel, E., Wiig, E., & Secord, W. (2003). *Clinical Evaluation of Language Fundamentals, Fourth Edition* (CELF-4). San Antonio, TX: NCS Pearson PsychCorp

Wallace, G., & Hammill, D, (2002) *Comprehensive Receptive and Expressive Vocabulary Test, Second Edition* (CREVT-2). Austin, TX: Pro-Ed.

Williams, K. (2007). *Expressive Vocabulary Test, Second Edition* (EVT-2). Minneapolis, MN: NCS Pearson.

.

Appendix B

Form Used for Reviewing the Language Tests

Criterion	Specific operational definitions based on McCauley & Swisher (1984) except for criterion 11	Met?	Page #	Comments/questions
1 DESCRIPTION OF THE STANDARDIZATION SAMPLE. The test manual should clearly define the standardization sample so that the test user can examine its appropriateness for a particular test taker (APA, 1974, pp. 20-21; Weiner & Hoock, 1973).	**"To pass this criterion, the test needed to give three pieces of information considered important for speech and language testing:** **(a) the normative sample's geographic residency,**			
	(b) socioeconomic status, and			
	(c) the "normalcy" of subjects in the sample, including the number of individuals excluded because they exhibited nonnormal language or nonnormal general development."			
2. SAMPLE SIZE For each subgroup examined during the standardization of the test, an adequate sample size should be used (APA, 1974, pp. 27-28, 37).	**"To pass the criterion, the test needed to have subgroups with a sample size of 100 or more."**			
3. CONTENT VALIDITY – ITEM	**"To pass the criterion, the test manual needed to**			

ANALYSIS. The reliability and validity of the test should be promoted through the use of systematic item analysis during item construction and selection (Anastasi, 1976, p. 198).	**report evidence that quantitative methods were used to study and control item difficulty, item validity, or both."**			
4. MEANS AND STANARD DEVIATIONS. Measures of the central tendency and variability of test scores should be reported in the manual for relevant subgroups examined during the objective evaluation of the test (APA, 1974, p. 22).	**"To pass this criterion, both the mean and standard deviation had to be given for the total raw scores of all relevant subgroups."**			
5. CONCURRENT VALIDITY. Evidence of concurrent validity should be supplied in the test manual (APA, 1974, pp. 26-27).	**"To pass this criterion, the test manual needed to provide empirical evidence that categorizations of children as normal or impaired obtained using the test agree closely with categorizations obtained by other methods that can be considered valid, for example, clinician judgments or scores on other validated tests."**			

6. PREDICTIVE VALIDITY. Evidence of predictive validity should be supplied in the test manual (APA, 1974, pp. 26-27).	**"To pass this criterion, a test manual needed to include empirical evidence that it could be used to predict later performance on another, valid criterion of the speech or language behavior addressed by the test in question."**			

7.TEST-RETEST RELIABILITY. An estimate of test-retest reliability for relevant subgroups should be supplied in the test manual (APA, 1974, pp. 50, 54).	**To pass this criterion, the test manual needed to supply empirical evidence of test retest reliability, including a correlation coefficient of .90 or better (Salvia & Ysseldyke, 1981, p. 98)**			
	that was statistically significant at or beyond the .05 level (Anastasi, 1976, pp. 108-109).			
8. INTEREXAMINER RELIABILITY. "Empirical evidence of interexaminer reliability should be given in the test manual (APA, 1974, p. 50)"	**"Evidence of interexaminer reliability that included a correlation coefficient of .90 or better (Salvia & Ysseldyke, 1981, p. 98)**			
	that was statistically significant at or beyond the .05 level (Anastasi, 1976, pp. 108-109)."			
9. DESCRIPTION OF ADMIN. PROCEDURES. "Test administration procedures should be described in sufficient detail to enable the test user to duplicate the administration and scoring	**"The test manual needed to provide sufficient description so that, after reading the test manual, the reviewer believed she could administer and score the test without grave doubts about correct procedures."**			

procedures used during test standardization (APA, 1974, p. 18).”				
10.TESTER QUALIFICATIONS. “The test manual should supply information about the special qualifications required of the test administrator or scorer (APA, 1974, p.15; Salvia & Ysseldyke, 1981, p.18).”	**General requirements (e.g., specific degree or years of experience)**			
	Specific requirements (e.g., number of times to have practiced administration; specific training offered)			
11. CONSTRUCT VALIDITY. **(Operational definition based on McCauley & Strand, 2008, p.84).**	**Any of the following were needed to meet the operational definition:** **(a) Evidence from a factor analytic study confirming expectations of the test’s internal structure**			
	(b) Evidence that test performance improves with age			
	(c) Evidence that groups that were predicted to differ in test performance actually do so.			
	In addition, evidence needed to be obtained within a study in which statistical methods [were described]			

	and participants were described.			